Stat 201: Introduction to Statistics

Standard 2: Describing an Experiment

Chapter One

Summaries

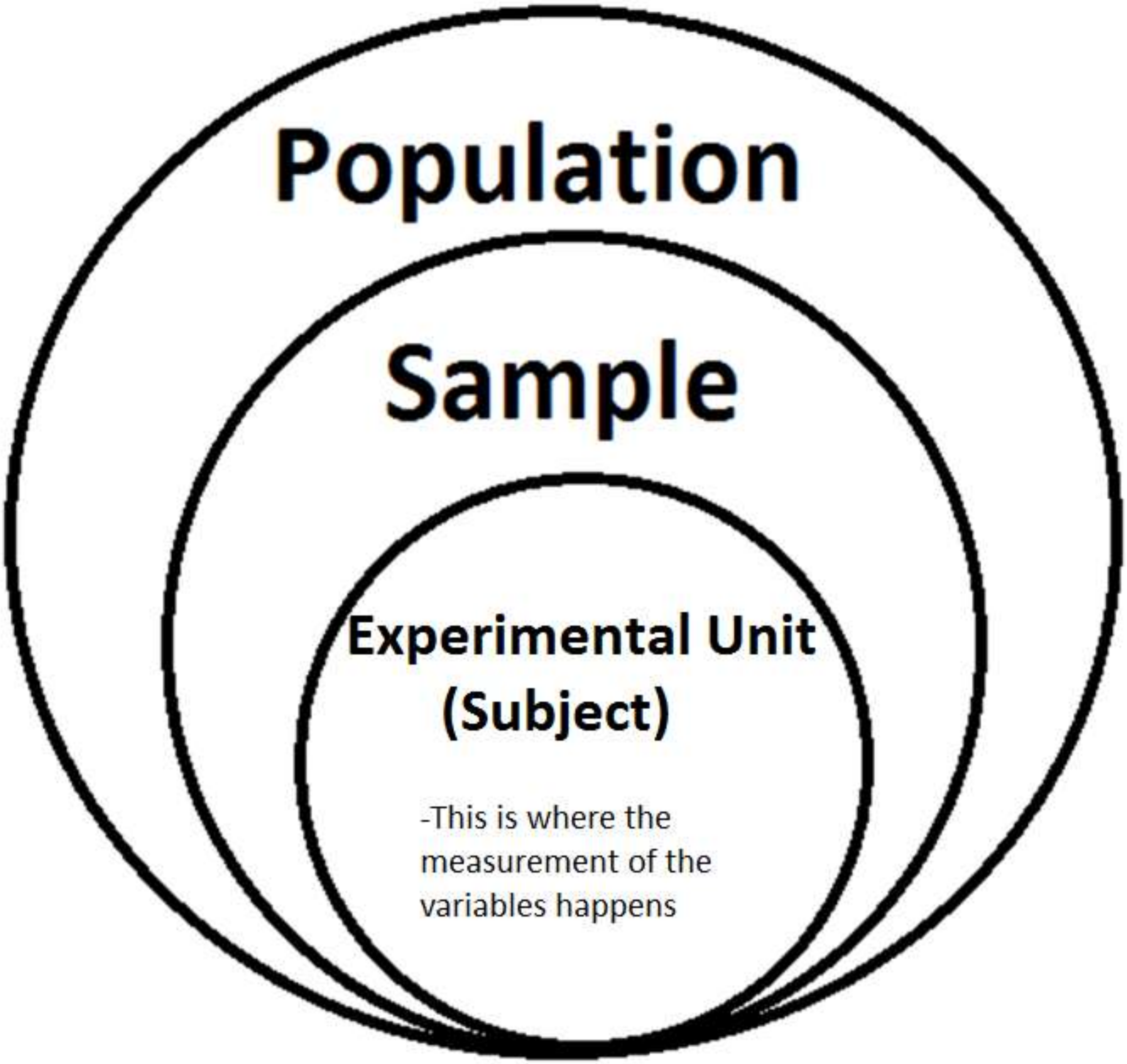
Definitions

- **Subject:** entities that we measure in a study
  - People, schools, the person or thing we look at
- **Population:** the set of all subjects of interest
  - US population, schools in SC, the group we look at
- **Variable:** any characteristic that is observed for the subject
  - Height, class size, whatever we're measuring
- **Sample:** the set of subjects that we have data for
  - A subset of the population for which we know the variable

Numerical Calculations

- **Statistic**: numerical summary of a sample
  - Mean($\bar{x}$), proportion($\hat{p}$),median, mode, standard deviation(s), variance(s^2), Q1, Q3, IQR, etc.
  - Statistic starts with an ‘s’ so it’s talking about the sample

- **Parameter**: numerical summary of a population
  - Mean(μ_x), proportion(ρ), median, mode, standard deviation(σ), variance(σ^2), Q1, Q3, IQR, etc.
  - Parameter starts with a ‘p’ so it’s talking about the population



Population

Sample

Experimental Unit (Subject)

-This is where the measurement of the variables happens

Observational Versus Designed

- An **observational study** measures the response variable without attempting to influence the value of either the response or explanatory variables.
- A **designed study** occurs when a researcher assigns the individuals or subjects into groups and intentionally affects their explanatory variables (think treatments)

Types of Sampling

Census	Collect data for every individual in the population (US Census)
Judgment	Collect a sample that an expert thinks is representative (lab one)
Systematic	Use a method (every 5th subject)
Convenience	Collect the sample that is easiest to access (this class)
Volunteer	Subjects choose to participate (medical experiments)
Simple Random Sample	The sample is chosen in such a way that every subject is equally likely to be selected for the study
Stratified Sample	Obtained by separating the population into non-overlapping groups called strata and taking a **SRS** from each strata individually
Cluster sample	Obtained by selecting all individuals or subjects within a randomly selected group

Types of Bias

Bias	When the results of a sample are not representative of a population
Sampling Bias	The sampling technique favored one part of the population over the other
Under coverage	When our sample doesn't follow the same proportions as the population
Nonresponse Bias	Occurs when individuals that respond to the survey answer differently than the potential answers of those who did not answer.
Response Bias	Occurs when the answers on a survey do not reflect the true feeling of the respondent.
Measurement Error Bias	Occurs when we have inaccurate data because of tools used.
Simple Random Sample	The sample is chosen in such a way that every subject is equally likely to be selected for the study

Describing an Experiment

Experiment	A controlled study conducted to determine the effect of varying one or more explanatory variables or factors has on a response.
Treatment	Any combination of the values of the factors that may affect the outcome
Experimental Unit	An individual, subject or thing to which a treatment is applied
Control Group	The baseline treatment that can be used for comparisons. This group is usually given a placebo.
Placebo	A non treatment such as saline or sugar tablets.

Types of Experiments

Completely Randomized Design	When each experimental unit is randomly assigned to a treatment.
Matched-pairs design	When experimental units are paired up based on some similarity before the treatment to test the difference after
Randomized block design	An experiment that uses the two methods below: blocks and blocking
Blocking	Grouping together similar experimental units and randomly assigning them within each group to a treatment
Block	Each group of similar experimental units

Blinding

Blinding	Refers to the hiding of which treatment is the real treatment and which is the placebo
Single Blind	When the recipient of the treatment doesn't know which they are getting
Double Blind	Is the single blind where those giving the treatment don't know either

Confounding and Lurking Variables

- **Confounding** in a study occurs when the effects of two or more explanatory variables are not separated
- This is often caused by a **lurking variable** which was not considered in the study but affects the response variable
- **Examples: http://www.tylervigen.com/**

Walkthrough

Example

- The 2012 South Carolina Republican Primary was held on January 21st. Newt Gingrich ended up winning the primary with 244,065 of 603,770 votes, 40.42% of South Carolina Primary voters.
  - **Population:** South Carolina Voters
  - **Sample:** Residents that came out to vote
  - **Experimental Unit(Subject):** Individual Voters
  - **Variable:** Which candidate they prefer
  - **Statistic:** $\hat{p} = \frac{244{,}065}{603{,}770} = .4042 = 40.42\%$

Example

- **Population:** South Carolina Voters
- **Sample:** Residents that came out to vote
- **Experimental Unit(Subject):** Individual Voters
- **Variable:** Which candidate they prefer
- **Statistic:** $\hat{p} = \frac{244{,}065}{603{,}770} = .4042 = 40.42\%$
  - The statistic 40.42%, a proportion, gives us information that suggests Newt Gingrich was preferred to other presidential hopefuls for all South Carolina Voters based off of those that voted
    - We will talk about how to find a reliability measure later in the semester

Example 2

- A ’93 survey of 4,977 found that 3.5% of households surveyed had used a gun for “protection” in the last year
  - **Population:** American households
  - **Sample:** 4,977 households sampled
  - **Experimental Unit(Subject):** Individual households
  - **Variable:** did they’ve used a gun for “protection?”
  - **Statistic:** $\hat{p} = .035 = 3.5\%$

Example 2

- **Population:** American households
- **Sample:** 4,977 households sampled
- **Experimental Unit(Subject):** Individual households
- **Variable:** Which candidate they prefer
- **Statistic:** $\hat{p} = .035 = 3.5\%$
  - – The statistic 3.5%, a proportion, can be applied to the population to give us an idea about the number of households that use their gun for protection in 1992: 3.5% of the population of 256.5M is 8,977,500
    - We will talk about how to find a reliability measure later in the semester

Example 3

- A poll surveyed 1,772 registered voters, 92 percent of which supported background checks, the Quinnipiac University telephone poll showed with a margin of error of plus or minus 2.3 percentage points
  - **Population:** Registered voters
  - **Sample:** 1,772 registered voters sampled
  - **Experimental Unit(Subject):** Individual registered voters
  - **Variable:** did they support background checks
  - **Statistic:** $\hat{p} = .92 = 92\%$

Example 3

- **Population:** Registered voters
- **Sample:** 1,772 registered voters sampled
- **Experimental Unit(Subject):** Individual registered voters
- **Variable:** did they support background checks
- **Statistic:** $\hat{p} = .92 = 92\%$
  - The statistic 92%, a proportion, can be applied to the population to give us an idea about the number of registered voters that support background checks.
  - **The margin of error** of 2.3% gives us an idea about the possible error in applying this to the population. In fact, we can expect between 89.7% to 94.3% of the population to support background checks.

Example 4

- CNBC reports that the average price of a gallon of gas at United States gas stations was $2.63 on Tuesday May 5th, 2015 according to reported prices on GasBuddy.com
  - **Population:** United States gas stations
  - **Sample:** United States gas stations reported on GasBuddy.com
  - **Experimental Unit(Subject):** Individual Stations
  - **Variable:** Price of a gallon of gas
  - **Statistic:** $\bar{x} = 2.63$

Example 4

- **Population:** United States gas stations
- **Sample:** United States gas stations reported on GasBuddy.com
- **Experimental Unit(Subject):** Individual Stations
- **Variable:** Price of a gallon of gas
- **Statistic:** $\bar{x} = 2.63$
  - The statistic \$2.63, a mean, can be applied to the population to give us an idea about the average price of gas at all stations in the United States.
    - We will talk about how to find a reliability measure later in the semester

Example 5

- The Wall Street Journal reports that on average, Panthers fans make 6.6 mistakes per 100 words placing them 19th among other teams, ages ahead of my poor Patriots. These results are based on a sample of 150 comments on each teams' website.
  - **Population:** All Panthers fans
  - **Sample:** Panthers fans that wrote 150 selected comments
  - **Experimental Unit(Subject):** Individual Panthers fans
  - **Variable:** Mean number of grammar mistakes per 100 words
  - **Statistic:** $\bar{x} = 6.6$

Example 5

- **Population:** All Panthers fans
- **Sample:** Panthers fans that wrote 150 selected comments
- **Experimental Unit(Subject):** Individual Panthers fans
- **Variable:** Mean number of grammar mistakes per 100 words
- **Statistic:** $\bar{x} = 6.6$
  - The statistic 6.6 mistakes per 100 words, a mean, can be applied to the population to give us an idea about the average grammar capabilities of Panthers fans.
    - We will talk about how to find a reliability measure later in the semester

Example 6

- Even though Pluto is no longer a full-fledged planet after it was downgraded to a dwarf planet it's still of interest. The gravitational approach allows scientists to use the gravitational approach to estimate Pluto's diameter at 1,471 miles, plus or minus five miles, by taking repeated measurements.
  - **Population:** All possible measurements
    - For now, we assume the population mean is the actual diameter
  - **Sample:** The collection of measurements taken
  - **Experimental Unit(Subject):** Each individual measurement
  - **Variable:** Diameter measurement of Pluto
  - **Statistic:** $\bar{x} = 1{,}471$

Example 6

- **Population:** All possible measurements
- **Sample:** The collection of measurements taken
- **Experimental Unit(Subject):** Each individual measurement
- **Variable:** Diameter measurement of Pluto
- **Statistic:** $\bar{x} = 1{,}471$
  - The statistic 1,471, a mean, can be applied to the population to give us an idea about the actual diameter of Pluto
  - The margin of error of 5 miles gives us an idea about the possible error in applying this to the population. In fact, we can expect the actual diameter to be between 1,467 to 1,475 miles.

Where Does Our Data Come From?

- We often hear reports about polls, research etc. on the news and the sources often a footnote are very important!
- We consider the following:
  1. Published Data
  2. Data from Designed Experiments
  3. Data from an Observational Study

Published Data

- This is just as it sounds. Many people have already completed designed experiments and observational studies and, thanks to the internet, have made their datasets publicly available.
  - https://www.kaggle.com/
    - Let's companies and researchers post datasets for other researchers to work with in hopes of finding new patterns
  - http://www.pewresearch.org/data/download-datasets/
    - PEW posts many of their datasets for secondary analysis
  - https://aws.amazon.com/datasets
    - Even Amazon posts public datasets from many categories

Observational Versus Designed

- An **observational study** measures the response variable without attempting to influence the value of either the response or explanatory variables.
- A **designed study** occurs when a researcher assigns the individuals or subjects into groups and intentionally affects their explanatory variables (think treatments)

Observational Versus Designed

- **A point of confusion:**
  - All experiments are designed in the sense of the English word in that someone plans them out. When I ask whether or not an experiment is designed or not – use the definition on the previous slide

Example

- C. Myrray Parkes headed a study of 4,486 men of 55 years of age and older who had their wives die in 1957. For up to nine years, these widowers were tracked and 213 died during the first six months – that’s about 5%.
- This experiment is a **observational study**. C Myrray Parkes didn’t murder 4,486 women in 1957 just to do this study.

Example

- Note: It was found that 40% above the expected rate for married men of the same age died during the first six months of bereavement. Thereafter, the mortality rate fell gradually to that of married men and remained at the same level.

Example 2

- Herbet Benson, MD headed a study in 2005 to see if intercessory prayer influenced recovery from bypass surgery. There were three groups in the study:
  1. Those being prayed for that didn't know
  2. Those being prayed for that did know
  3. Those not being prayed for
- This is a **designed study** because the researchers assigned different patients to different groups; they controlled who was prayed for and who wasn't instead of just observing and asking the families whether or not they had friends and families praying for the patient.

Example 2

Note: Intercessory prayer itself had no effect on complication-free recovery but knowing they were receiving intercessory prayer was associated with a higher incidence of complications.

Where We Get Our Sample is Important!

- Regardless of which of the three options we get our data from when delivering descriptive or inferential statistics it is paramount that we have a **representative sample**
- The idea is that the sample we use should accurately portray the description of the population we’re trying to talk about; we don’t want to compare apples to oranges!

Example

- Suppose we want to measure the age of lung cancer instances in smokers and non-smokers and consider the sample made up of smokers in their 80’s and 90’s and non smokers of any age.
- Overall, the average age of a lung cancer patient is about 70 years old. With the proposed sample above the minimum average age of lung cancer instances is 80, well above the overall average; this could lead to us saying that cigarettes are good for you! We note, however, that people start smoking much earlier than their 80’s and 90’s so our sample is **not representative.** We say this type of ‘non-representative’ sample suffers from **selection bias**.

Example

- Suppose we want to measure the age of lung cancer instances in smokers and non-smokers and consider a sample of cigarette smokers and non-cigarette smokers both with wide age ranges and backgrounds that volunteered to be part of the study.
- With the proposed sample above we have comparable experimental units (or subjects) in both groups and because of their wide age ranges and backgrounds we can argue that this sample is **representative** of the population and that the results can be generalized to the public.

Major Components to Statistics

- 1) **Design of Study**
  - What question are we answering?
  - What do we need to look at?
- 2) **Descriptive Statistics**
  - What summary can help us answer the question?
- 3) **Inferential Statistics**
  - Can we predict or draw conclusions based on the two other components?

Design of the Study

- What is the research question?
- What is the **population** of interest?
- What is the **variable** of interest?
- How will the data be collected?
- How will the **sample** be selected?

- Essentially, what's the best way of going about using statistics to solve your problem?

Design of the Study: Sample Selection

- **Census** – collect data for every individual in the population
  - Problem: time, money, usually impossible
  - Example: Country Census Data, World Bank Data
- **Judgment** – collect a sample that an expert thinks is representative
  - Problem: there may be some bias
  - Example: Surveying land for contamination (Lab 1)

Design of the Study: Sample Selection 2

- **Convenience** – Collect the sample that is easiest to access
  - Problem: sample will be bias
  - Example: If I were to use this class instead of randomly sampling the entire university
- **Volunteer** – Subjects choose to participate
  - Problem: sample will likely be biased
  - Example: Medical experiments to test medications

Design of the Study: Sample Selection 3

- **Systematic** – Use a method (every 5th subject)
  - Example: Check every fifth item produced
  - Problem: there may be a system we don’t know
    - Maybe every fifth item was made by the same machine

Design of the Study: Sample Selection 4

- **Stratified Sample** – obtained by separating the population into non-overlapping groups called strata and taking a **simple random sample** from each strata individually

Design of the Study: Sample Selection 5

- **Cluster sample** – obtained by selecting all individuals or subjects within a randomly selected group. This is similar to stratified but the subjects are already grouped
  - Taking a random sample of classes at Florence-Darlington Tech instead of a random sample of students is an example of this

Desired Sample Selection

- **Simple Random Sample** – the sample is chosen in such a way that every subject is equally likely to be selected for the study
  - We prefer this method above all else
  - Problem: Sometimes this isn’t feasible
    - We don’t always have access to every subject in the population
    - It’s not always the case that everyone is willing to participate in the study

Bias

- **Bias** is when the results of a sample are not representative of a population

Sources of Bias

- **Sampling or selection bias** means that the sampling technique favored one part of the population over the other, i.e. parts of the population have no chance of being selected for the sample
  - Experiments using a convenience sample usually suffers from sampling bias; ; consider if I took students at USC as a ‘random’ sample of Americans because I have convenient access to students as a GA. In this case I would be leaving out all non-students from the sample leading to sampling bias
  - **Undercoverage** is when our sample doesn’t follow the same proportions as the population

Sources of Bias

- **Nonresponse Bias** occurs when individuals that respond to the survey answer differently than the potential answers of those who did not answer.
  - How many times do you participate in the optional survey at when using the web or calling customer service? Maybe only the people that are angry respond; this leaves all the satisfied customers out of the sample and reflects higher dissatisfaction than it might have otherwise.

Sources of Bias

- **Response Bias** occurs when the answers on a survey do not reflect the true feeling of the respondent.
  - How often do you really take your time to fill out teacher evaluations?
  - Maybe a respondent will answer dishonestly about personal questions.

Sources of Bias

- **Measurement Error Bias** occurs when we have inaccurate data
  - Often survey data comes from interviewing individuals.
    - Sometimes the content of a survey is sensitive and participants feel the need to lie or omit details in their answer, interviewers can ask leading or confusing questions, or interviewers could have language difficulties
    - Sometimes our tools fail us and can produce faulty measurements
    - Sometimes we don’t use the right or ‘best’ tool for the job

Describing an Experiment

- An **experiment** is a controlled study conducted to determine the effect of varying one or more explanatory variables or factors has on a response.
    - Finding how explanatory variables explain the response
- A **treatment** is any combination of the values of the factors

Describing an Experiment 2

- An **experiment unit** is an individual, subject or thing to which a treatment is applied
- A **control group** serves as the baseline treatment that can be used for comparisons
- A **placebo** is a non treatment such as saline or sugar tablets.

Describing an Experiment 3

- **Blinding** refers to the hiding of which treatment is the real treatment and which is the placebo
- **Single Blind** is when the recipient of the treatment doesn’t know which they are getting
- **Double Blind** is the single blind where those giving the treatment don’t know either.

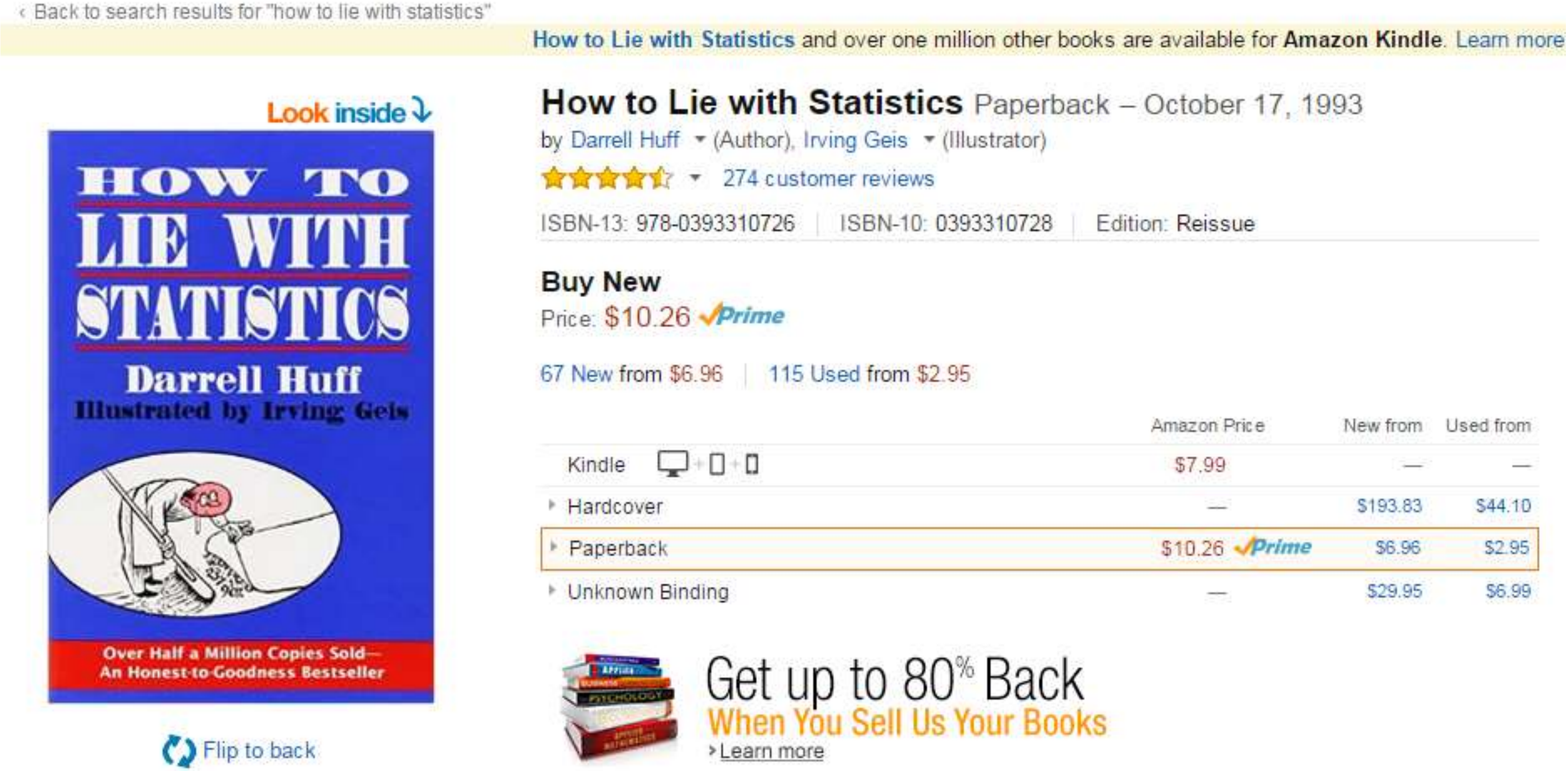
Experimental Designs

- A **completely randomized design** is when each experimental unit is randomly assigned to a treatment.

- A **Matched-pairs design** is when experimental units are paired up based on some similarity before the treatment to test the difference after

Experimental Designs

- Grouping together similar experimental units and randomly assigning the experimental units within each group to a treatment is called **blocking**
- Each group of homogeneous individuals is called a **block**
- An experiment that uses these methods is said to use **randomized block design**



‹ Back to search results for "how to lie with statistics"

How to Lie with Statistics and over one million other books are available for **Amazon Kindle**. Learn more

Look inside
HOW TO LIE WITH STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller
Flip to back

How to Lie with Statistics Paperback – October 17, 1993

by Darrell Huff (Author), Irving Geis (Illustrator)

★★★★☆ 274 customer reviews

ISBN-13: 978-0393310726 | ISBN-10: 0393310728 | Edition: Reissue

Buy New

Price: $10.26 ✓Prime

67 New from $6.96 | 115 Used from $2.95

	Amazon Price	New from	Used from
Kindle	$7.99	—	—
Hardcover	—	$193.83	$44.10
Paperback	$10.26 ✓Prime	$6.96	$2.95
Unknown Binding	—	$29.95	$6.99

Confounding and Lurking Variables

- **Confounding** in a study occurs when the effects of two or more explanatory variables are not separated

- This is often caused by a **lurking variable** which was not considered in the study but affects the response variable
- **Examples: http://www.tylervigen.com/**

Confounding and Lurking Variables

- It’s hot out
  - When it’s hot where do you go?
    - The pool
  - When it’s hot what do you eat?
    - Ice cream
- If we had a hot summer there would be more swimming and, thus, more drowning deaths. If someone wasn’t careful it might look like ice cream sales cause drowning deaths – this is **confounding** and in this case the temperature would be the **lurking variable.**